



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

200
1822 - 2022

Descriptive Statistics

Man Geni Addysg Uwch yng Nghymru
Where Higher Education Began in Wales

www.uwtsd.ac.uk

Objectives

- Mean
- Median
- Mode
- Range
- Standard Deviation (σ)
- Variance (σ^2)

Descriptive Statistics

Frequency distributions - graphical view of data

Drawback: Not enough or right kind of information

“What is the average GRE score of this class?”

“How much money does the average football player make?”

- on **average**, the annual growth rate of stocks during the past 10 years has exceeded by 20 points.
- You might stop smoking because, on **average**, non-smokers can expect to live longer than heavy smokers (by as much as 10 years, according to some researchers).
- You might strengthen your resolve to graduate from college upon hearing that, on **average**, the lifetime earnings of college graduates are almost double those of high school graduates.
- To do this we need to look at
 - Measures of central tendency – Mean, Median, Mode
 - Measures of variability- Range, Variance(SD^2), Standard

Deviation (SD)

Measure of central tendency- *Mean*

$$\bar{X} = \frac{\sum X}{N}$$

Where,

X is the set of scores and

N is the number of scores

Mean is denoted by $\bar{X}$

Example

Maths Score

14

19

13

9

13

$$\bar{X} = \frac{\sum X}{N} = \frac{68}{5} = 13.6$$

Reading Score

25

37

26

20

19

$$\bar{Y} = \frac{\sum Y}{N} = \frac{127}{5} = 25.4$$

The ***Median*** – Middle value

- divides the frequency distribution in half

Example 1: odd set of scores

113, 95, 94, 133, 112

- Sorting : 94, 95, 112, 113, 133
- Middle value: 112

Example 2: even set of scores

113, 95, 94, 133, 112, 102

Sorting: 94, 95, 102, 112, 113, 133

Mean of the middle two values:

$$(102 + 112) / 2 = 107$$

Mode - most frequently occurring score in a set

Example 1:

Given a set of scores

110, 105, 100, 100, 100, 100, 99, 98

100 appears 4 times

Example 2:

Given a set of scores

110, 105, 105, 103, 100, 100, 99, 98

105 and 100 both appear twice

Multimodal distributions are also possible

Bimodal

Exercise

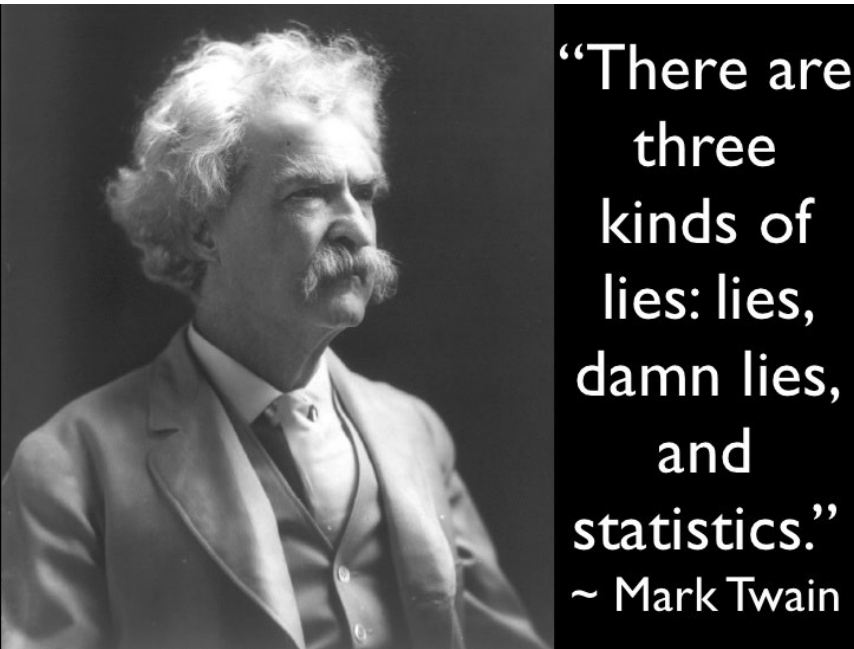
- Determine the mean, median and mode(s) of the midterm examination dataset (20 students)

	Mean		median	mode	
23	63.16667		67	78	
45					
56					
78					
78					
99					



Microsoft Excel
97-2003 Workshee

Selecting a measure of central tendency



"There are
three
kinds of
lies: lies,
damn lies,
and
statistics."
~ Mark Twain

Symmetrical unimodal
distribution

Mean = = Median == Mode

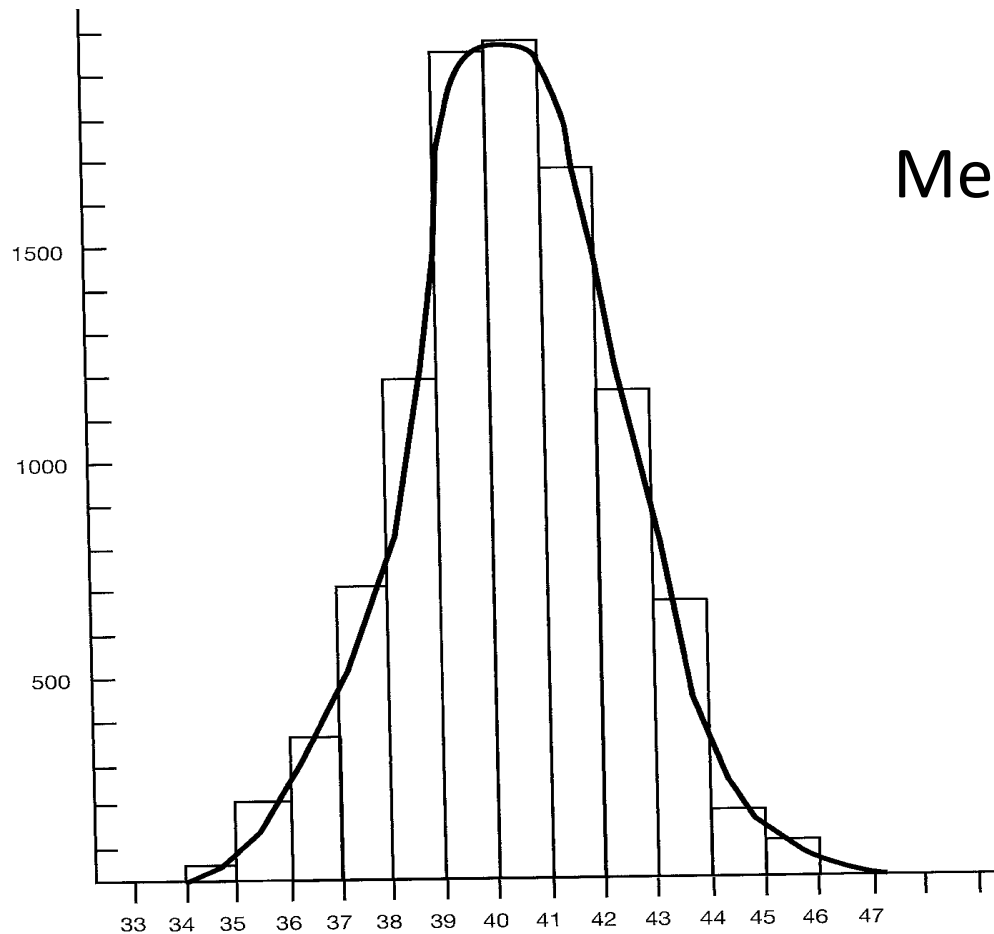
Mean is the 'balance point'
of a distribution

Skewed Dataset (unbalanced Dataset)



Find Median and not Mean

The Normal curve



Mean-best Choice

FIGURE 4-3 Chest sizes of 5,738 soldiers.

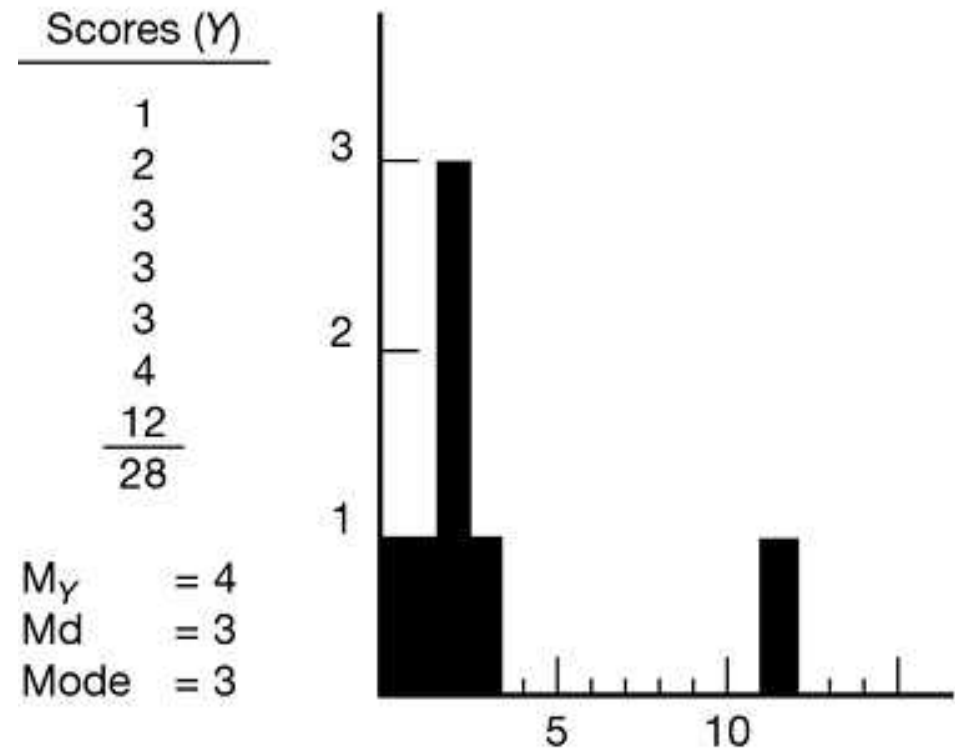
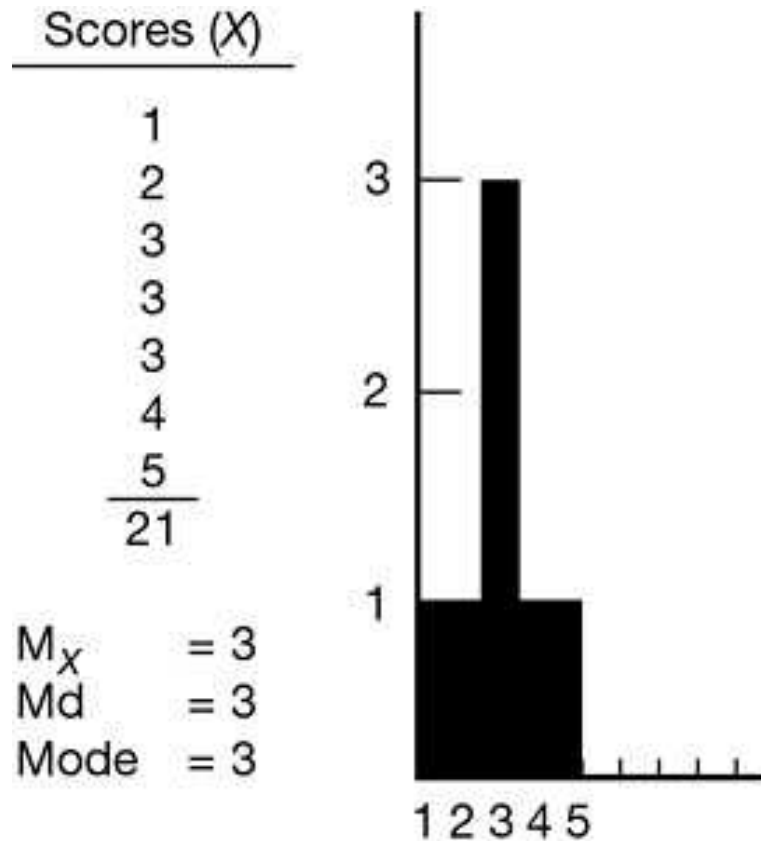
Income Comparison of Football Team's Player

Consider a professional football team with a mix of players. Most of the players earn standard salaries that are typical for professional athletes, but the team also includes a few star players with exceptionally high salaries due to their outstanding performance and market value. Annual income:

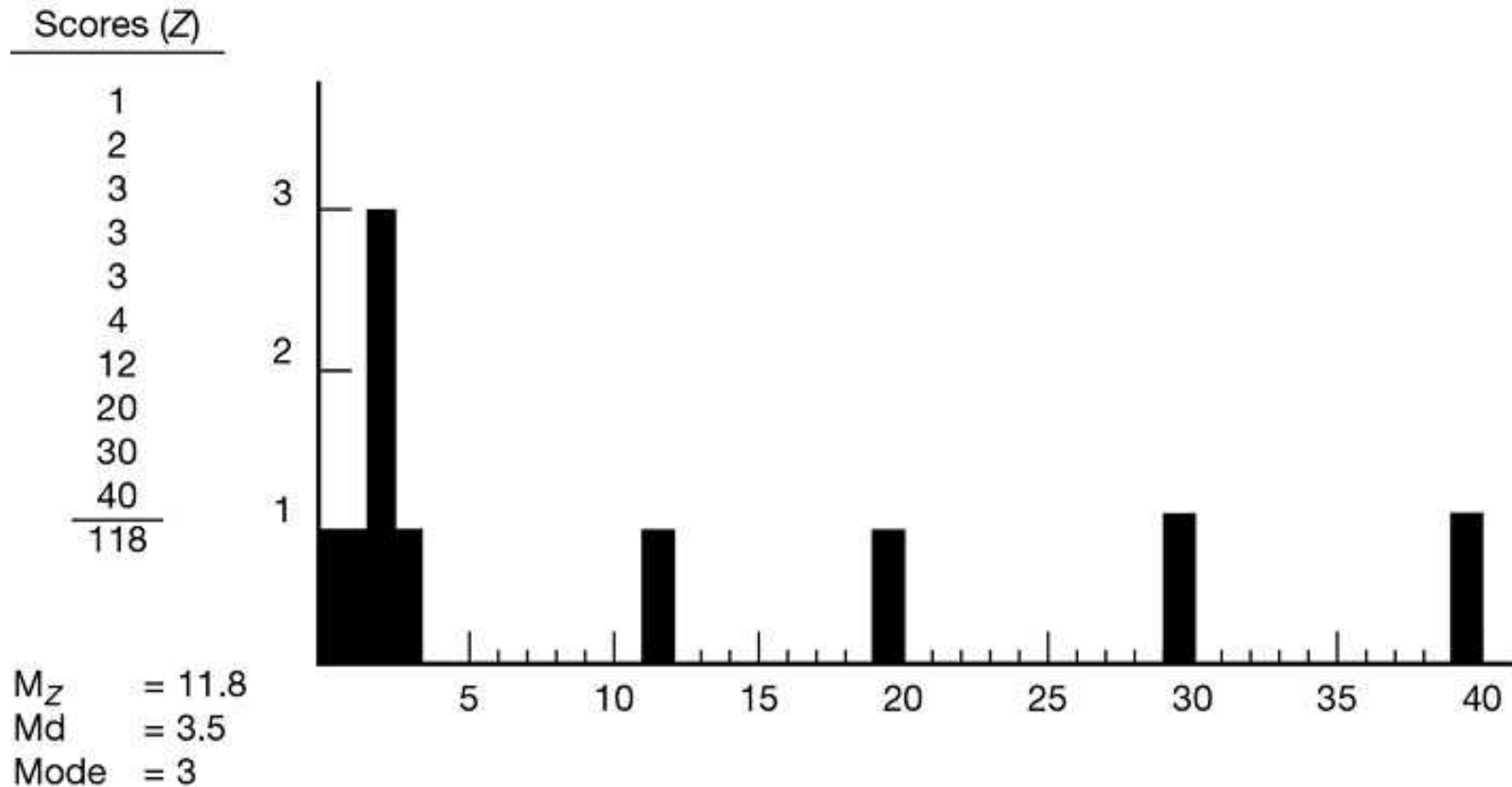
250,000	25,000
260,000	30,000
270,000	35,000
280,000	40,000
280,000	45,000
1,000,000	50,000

Alternative Approach: In this scenario, using the median salary might offer a clearer picture of what the typical player earns. Represent the salary at the midpoint when arranged in ascending order, not affected by the extreme salaries of the star players.

Variation with Skewness



Variation with Skewness



Measures of Variability

- Range
- Standard Deviation (σ)
- Variance (σ^2)

represents the spread or dispersion of data points around the central tendency

Range

$$\text{Range} = H - L$$

Where H- highest score & L – Lowest score

Classroom X Exam Score

160

161

130

100

70

40

$$\begin{aligned}\text{Range} &= 160 - 40 \\ &= 120\end{aligned}$$

Skewness not considered

In Excel → functions like **MAX** and **MIN** used

Standard Deviation

- The Standard Deviation is a measure of how spread out numbers are.
- Its symbol is σ (the greek letter sigma)
- The formula is easy:
It is the square root of the Variance.

So now you ask, "What is the Variance?"

Variance:

defined as the mean of the squared deviation scores.

$$\bar{X} = \frac{\sum X}{N} = \frac{25}{5} = 5 \quad \sigma^2 = \frac{\sum (X - \bar{X})^2}{N} = \frac{40}{5} = 8$$

Scores (X)	$\bar{X}$	$(X - \bar{X})$	$(X - \bar{X})^2$
9	9 - 5	4	16
7	7 - 5	2	4
5	5 - 5	0	0
3	3 - 5	-2	4
1	1 - 5	-4	16
$\sum X = 25$			$\sum (X - \bar{X})^2 = 40$

Variance

Population

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

X - The Value in the data distribution

μ - The population Mean

N - Total Number of Observations

Sample

$$\sigma^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$$

X - The Value in the data distribution

$\bar{X}$ - The Sample Mean

n - Total Number of Observations

In excel it can be differentiated by `var.p( )` and `var.s( )`

Standard Deviation

Population

$$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}}$$

X - The Value in the data distribution

μ - The population Mean

N - Total Number of Observations

Sample

$$s = \sqrt{\frac{\sum(X - \bar{x})^2}{n - 1}}$$

X - The Value in the data distribution

$\bar{x}$ - The Sample Mean

n - Total Number of Observations

Standard Deviation

Population

$$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}}$$

X - The Value in the data distribution

μ - The population Mean

N - Total Number of Observations

Standard Deviation (S.D.)

Scores (X)

9

7

5

3

1

Variance $SD^2 (\sigma^2) = 10$

Standard Deviation $SD(\sigma) = \sqrt{10}$
 $= 3.16$

Exercises

- **Exam Performance:** A school administrator wants to compare the academic performance of students from different classes to identify any significant differences. What measure of central tendency would be best for the school administrator to use in a standardized test? Explain.

Solution:

Measure of Central Tendency: Mean

Comparing the average scores obtained by each class in a standardized test would be best using the mean

Exercise(contd..)

Traffic Analysis: A city planner wants to understand the typical daily traffic congestion levels on different routes within the city. They seek a measure of central tendency to represent the average traffic volume for each route to assist in traffic management strategies.

Solution:

Measure of Central Tendency: Mode or Median

Analyzing traffic congestion levels might benefit from either the mode (most common level of congestion) or the median traffic volume for each route. The mode represents the most frequent level of congestion, while the median provides a middle value, giving insight into typical traffic volumes

Exercise(contd..)

A researcher gathers IQ data for a large group of undergraduate students. The mean of his distribution of scores is 110, and the median is 130. What can he assume about the shape of his distribution?

Solution:

This discrepancy between the mean and median suggests that the distribution is likely **skewed to the left** (negatively skewed).

- In a negatively skewed distribution:
- The mean is smaller than the median.

Exercise

Find the mean, range, variance and standard deviations for each of the following groups of scores. Carry out all calculations to three decimal places and round them correct to two places

Classroom X Exam Score

160

100

80

69

69

AutoSave Off Worksheet in week 8 - Descriptive Statistics - Compatibility Mode Search (Alt+Q)

File Home Insert Draw Page Layout **Formulas** Data Review View Developer Help Acrobat

Insert Function AutoSum Recently Used Financial Logical Text Date & Time Lookup & Reference Math & Trig More Functions Name Manager Define Name Use in Formula Trace Precedents Trace Dependents Remove Arrows Show Formulas Error Checking Evaluate Formula

Function Library

Statistical > Engineering > Cube > Information > Compatibility > Web >

AVEDEV
AVERAGE
AVERAGEA
AVERAGEIF
AVERAGEIFS
BETA.DIST
BETA.INV
BINOM.DIST
BINOM.DIST.RANGE
BINOM.INV
CHISQ.DIST
CHISQ.DIST.RT
CHISQ.INV
CHISQ.INV.RT
CHISQ.TEST
CONFIDENCE.NORM
CONFIDENCE.T

	A	B	C	D	E	F	G	H	I
1		Mean		median	mode				
2	23	63.16667		67	78				
3	45								
4	56								
5	78								
6	78								
7	99								
8									
9		743.7667							
10									
11		619.8056							
12									
13									
14									
15	9	8	10		160	95.6	80	69	11
16	7				100				1
17	5				80				
18	3				69				
19	1				69				
20									
21									

Using Excel

D7 ✕ ✓ <i>fx</i>				
	A	B	C	D
1	Aptitude test scores			
2	160	Mean		95.6
3	100	Median		80
4	80	Mode		69
5	69	Variance		1165.04
6	69	Standard Deviation		34.13268
7				
8				

Summary

- All of today's material looked at descriptive statistics of univariate data
 - Measures of central tendency
 - Selecting a measure of central tendency
 - Measures of variability
 - Using Excel for statistics

References:

Chapter 5- *Statistics for the Terrified*.